SNR Equation Spec Sheet (Canonical)
Version: 1.0 • Status: Canonical • Window Size: 2000 words

Deterministic rubric for word-level Signal-to-Noise Ratio (SNR) measurement. Reproducible and auditable.

1) Tokenization

Input = raw text.
Normalize: replace curly quotes/apostrophes/dashes with straight equivalents.
Tokenizer (regex, word-level): `\b[\w']+\b` (captures words+digits, ignores punctuation).
All downstream rules operate on these tokens in order given below.

2) Classification Rules (exclusive precedence order)

A. Number check — if token.isdigit() → NOISE [timestamps/IDs/scaffolding].
B. Length check — if len(token) ≤ 2 → NOISE [near-zero information density].
C. Stopword check — if token.lower() ∈ English stopword set → NOISE [grammatical glue].
D. Filler (single-word) — if token.lower() ∈ {um, uh, like, just, really, actually, literally, okay, right, well, maybe, ...} → NOISE.
E. Filler (multi-word spans) — scan raw text for phrases (e.g., “you know”, “i mean”, “sort of”, “at the end of the day”); any word inside such spans → NOISE.
F. Default — otherwise → SIGNAL.

3) Counting

Maintain: signal_count, noise_count. Total = signal_count + noise_count.
(If a neutral class is later added, exclude neutrals from both signal and noise.)

4) Equations

```
SNR_ratio = Signal_Words / max(Noise_Words, 1)
SNR_dB    = 10 * log10(SNR_ratio)
Signal%   = Signal_Words / (Signal_Words + Noise_Words)
```

5) Windowed Analysis

Segment sequentially into equal windows of 2000 WORDS (not characters, not tokens of another tokenizer).
Compute the same metrics per window to produce drift profiles across the text.

6) Outputs (minimum artifacts)

summary.json — totals and top terms
windows.csv — per-window stats (index, start_word, end_word, Signal, Noise, SNR_ratio, SNR_dB, Signal%)
per_word.jsonl — one line per token: {index, word, lower, is_signal, reason}

7) Interpretation

SNR > 1 (positive dB) → signal outweighs noise (content-dense).
SNR ≈ 1 (~0 dB) → balanced.
SNR < 1 (negative dB) → noise outweighs signal (scaffolding/filler dominates).
Note: TPW (tokens-per-word) is diagnostic only; it is NOT a proxy for SNR.

8) Reproduction Requirements

Use the same tokenizer regex, the same 6-step precedence, identical equations, and fixed 2000-word windows.
Report totals AND window-level stats. Any deviation must be declared as a versioned fork.